Victor Jongeneel
is Director of the Office of Information Technology of the Ludwig Institute, and of the Swiss Institute of Bioinformatics.

Searching the expressed sequence tag (EST) databases: Panning for genes

C. Victor Jongeneel

Date received (in revised form): 22nd October 1999

Abstract

The genomes of living organisms contain many elements, including genes coding for proteins. The portions of the genes expressed as mature mRNA, collectively known as the transcriptome, represent only a small part of the genome. The expressed sequence tag (EST) databases contain an increasingly large part of the transcriptome of many species. For this reason, these databases are probably the most abundant source of new coding sequences available today. However, the raw data deposited in the EST databases are to a large extent unorganised, unannotated, redundant and of relatively low quality. This paper reviews some of the characteristics of the EST data, and the methods that can be used to find novel protein sequences within them. It also documents a collection of databases, software and web sites that can be useful to biologists interested in mining the EST databases over the Internet, or in establishing a local environment for such analyses.

Keywords: *database searching, gene discovery, expressed sequence tags, contig assembly, clustering*

Introduction

The ultimate goal of genome projects is to produce a complete and accurate sequence of the entire genetic material of a biological species. It was realised at an early stage that the part of the genome expressed as mRNA, often dubbed the *transcriptome*, contains much of the information of interest to biologists. In higher eukaryotes, the transcriptome represents only a small portion of the genome; in mammals, about 7 per cent of the genome is thought to be potentially expressed as mRNA, but the proportion in most differentiated tissues is much smaller. Experimentally, a crude snapshot of the transcriptome of a particular tissue or cell type can be obtained by producing a cDNA library of sufficiently high complexity, and sequencing a sufficiently large number of clones, to ensure that most of the information present in the library has been extracted. In order to keep the amount of work to manageable levels, and because much of the process has to be automated, single unverified reads are normally obtained from the 3' and 5' ends of each cDNA clone.

This approach, dubbed expressed sequence tag (EST) sequencing,[1,2] has been enormously successful in the framework of many genome projects. EST sequences contain at least partial sequences of most mRNAs present in the various tissues used for library construction. Therefore, they have been used intensively as a source of information for the discovery of new genes whose function can be tentatively deduced from their sequence, and experimentally verified. This review will focus on methods and tools used to detect the presence of new protein sequences of biological interest in the EST databases.

Because of the unique nature of EST data, the gene discovery process can be more complex than one may intuitively be led to believe. This paper discusses some of the issues to be considered in analysing EST data, and the practical steps one may take to deal with them.

THE DATA

The cDNA libraries from which EST sequences are extracted are usually constructed by traditional methods,

C. Victor Jongeneel,
Office of Information Technology,
Ludwig Institute for Cancer Research and
Swiss Institute of Bioinformatics,
chemin des Boveresses 155,
CH-1066 Epalinges,
Switzerland

Tel: +41 21 692 5991
Fax: +41 21 692 5945
E-mail: Victor.Jongeneel@licr.org

cDNA libraries

using oligo(dT) to prime first strand synthesis and various protocols to obtain the second strand. In almost all cases, the process produces 'oriented' clones, where the positions of the 5' and 3' ends of the cDNA relative to the vector are known in principle (although subject to some experimental error). Thus, two defined vector-based primers can be used to obtain a 3' and a 5' sequence from the same clone; depending on the length of the insert and the quality of the trace data, the sequences determined from the two ends may or may not overlap (Figure 1).

Quality issues

A single read is taken from each primer, and no effort is made to ensure that both reads from a given clone are of good quality. Current submission rules for the US National Center for Biotechnology Information (NCBI), which receives the bulk of EST data, require that the 'high-quality' part of submitted sequences meet minimal quality criteria, normally a calculated error rate of less than 1 per cent, corresponding to a 'phred score' of 20 or better. Phred is a base-calling program developed by Phil Green at the University of Washington, and used at most major sequencing centres.[3] However, these quality criteria are far from being met by all EST sequences currently in the public databases.

Digital differential display

Most of the cDNA libraries used for the generation of EST data (in particular those produced for the US National Cancer Institute's (NCI's) Cancer Gene Anatomy Project, CGAP), have been prepared to randomly sample the transcriptome of the tissue from which they were derived. In these libraries, the relative abundance of clones derived from a particular mRNA roughly reflects the abundance of this mRNA in the tissue from which the library was derived. This has the advantage of allowing 'digital differential display', ie the identification of genes that are more highly expressed in one sample than in another, based on the number of clones derived from that gene. The disadvantage is that cDNAs derived from genes expressed at a low level will be represented by a few clones at best. In some cases, the cDNA libraries have been 'normalised', meaning that some method (usually self-hybridisation) has

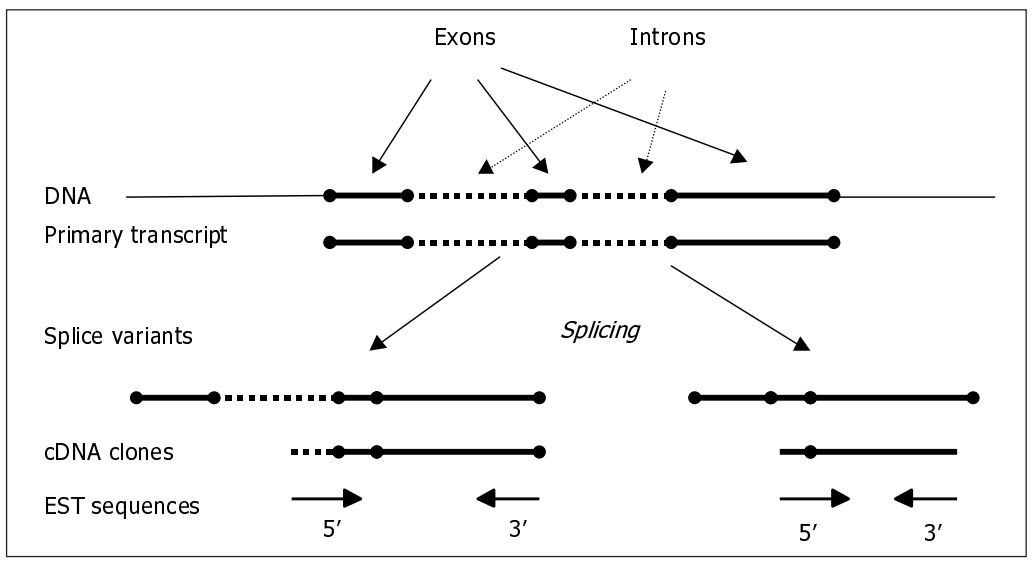

Figure 1: Relationship between EST sequences and RNA transcripts. The figure illustrates three major points : (1) The 3' EST sequences derived from individual clones representing the same gene will usually overlap. (2) The 5' sequences may overlap, but often do not. (3) Alternative or incomplete splicing can cause divergences within groups of ESTs derived from the same gene

Normalisation

been used to reduce the representation of highly expressed genes, and thus enhance the probability of finding clones derived from rarer mRNAs.[4]

These methodological considerations have several consequences for the nature and the quality of the sequence data found in the EST databases:

Representation

- Many sequences are derived from the 3' ends of mRNAs, and thus contain mostly information about 3' untranslated regions (UTRs) rather than the coding regions of genes. In fact, genes with very long 3' untranslated regions (UTRs) may be represented only by clones derived from these 3' UTRs, and no or very little coding region information may be found in the databases.

Contamination

- The average quality of the sequences is rather low; therefore, frameshift errors due to insertions and deletions, as well as artefactual stop codons, are quite common.

Clusters

- Genes that are highly expressed in the tissues from which libraries have been prepared will be represented in many EST sequences. Sequences derived from genes whose expression is low, or restricted to cell types that are underrepresented in the libraries, will be found in very few (if any) EST entries.

- Genes that are expressed only in tissues, cell types or developmental stages that were not used for preparing cDNA libraries will not be represented at all in the EST databases.

- There are a substantial number of sequences derived from partially spliced RNA species, which are often indistinguishable from *bona fide* splice variants. Chimeras, resulting from the artefactual ligation of unrelated cDNAs, are also common.

- The EST sequence collections are only as good as the libraries from which they were generated. There are many documented cases of contaminations by genomic DNA, by bacterial DNA, by cDNA from other species (ranging from fungi to mammals), by vector DNA, and by mitochondrial or ribosomal DNA. Many of these contaminants are still found in the public EST databases: for example, a stringent BLAST search against dbest (17th June 1999), using human 28S rRNA as a query, retrieved 3,176 hits representing contaminations of mammalian origin. Recent improvements in the quality control of EST sequences being deposited in GenBank should gradually reduce the frequency of easily detectable contaminants; however, the inclusion of genomic sequences and the erroneous annotation of species of origin are probably unavoidable problems.

EST CLUSTERING

Large-scale EST sequencing projects have generated many more sequences than there are expressed genes, or distinguishable mRNA species, in the organism under study. For example, there are currently over 1.5×10^6 public human EST sequences, derived from approximately 10^5 genes. Hence, there are many genes from which more than one EST was derived, and much of the sequence information in the EST databases is redundant. In order to establish a non-redundant catalogue of the genes represented in the EST collections, it is necessary to *cluster* EST sequences into groups that are likely to have been derived from the same gene, or even the same RNA species (to take into account splice variants). This is far from being a trivial exercise, and a discussion of clustering strategies is beyond the scope of this review (see, for example, refs 5–7).

The most enduring effort at EST clustering is the Unigene project of the NCBI.[8] Unigene is currently limited to

Unigene

four species: human, mouse, rat and zebra fish. In addition to EST data, it also includes mRNAs derived from known genes, and 'virtual' mRNAs deduced from the annotation of genomic sequences, culled from GenBank. In its rawest form, Unigene consists of a list of lists: each cluster is assigned a number (eg Hs.12345), and a list of accession numbers of ESTs and known mRNAs or gene transcripts belonging to the cluster. In addition, many useful annotations are added to each cluster: gene name (if known), similarities to known genes, chromosomal localisation, libraries of origin of the ESTs populating the cluster, tissue specificity of expression, etc. Unigene is also distributed as a subset of dbest, containing for each EST the name of the cluster to which it was assigned. This subset is not included in the BLAST searchable databases at NCBI; however, Unigene ESTs can be searched on the web site of the Swiss Institute of Bioinformatics (SIB), with links from the hits to Unigene cluster entry descriptions at NCBI.

N.B. the URLs for all the webservers discussed in this paper can be found in the Resources section at the end.

Gene indices

Ideally, Unigene would be a stable index of uniquely identifiable genes, where new ESTs would either be added to existing clusters or define new genes. Unfortunately, the state of the art in clustering does not allow this to be done yet: cluster numbers, and hence potential gene identifiers, change constantly as clusters are merged or split with each methodological improvement and update of the database. It is hoped that this situation will soon be resolved, following the recent announcement of the release of a stable human cDNA index collection by the National Institutes of Health (NIH). Nevertheless, identification of an EST as a member of a Unigene cluster greatly enhances the amount of information that can be gathered about the corresponding gene.

Clustering databases and software

There have been several other attempts at clustering ESTs. The STACK project at the South African National Bioinformatics Institute (SANBI) is releasing EST clusters that have also been assembled into contigs (see below). ESTs incorporated in STACK are sorted by tissue of origin before clustering. This reduces cluster size, and thus the complexity of the assembly problem but also precludes assembly of mRNAs whose corresponding ESTs were derived from libraries spanning multiple tissues.

The Institute of Genome Research (TIGR) has created unique gene indices of clustered and assembled ESTs,[9] available to academic institutions. The stringent approach adopted by TIGR has resulted in the production of larger numbers of high-quality contigs, where one gene may be represented in multiple entries. Several of the genome-oriented companies that have recently sprung up are reported to have assembled clusters and contigs representing most of the human transcriptome; however, only scientists at institutions willing to pay the very steep access fees can testify to the truth of this assertion.

The *ab initio* clustering of ESTs from a large collection is a very complex problem. There are a few public-domain software tools available to do it, including J. Parson's ICAtools[6] and JESAM suites (EBI), and the STACK_PACK suite distributed by SANBI. On the other hand, the generation of a cluster starting from a defined query sequence is a much easier task, and tools to perform such a task are available on a number of web sites (listed at the end of this paper).

EST CONTIG ASSEMBLY

A cluster of ESTs derived from the same gene can be extracted from Unigene, or from a hit list of an EST database search using an 'interesting' query (see below). In either case, it is worthwhile to try to derive a 'consensus sequence' from the ESTs in the cluster, and thus eliminate redundancy and reduce the error rate,

Trace data

while increasing the length of the deduced mRNA sequence. This problem is very similar to the generation of a contig from shotgun sequences, as is often performed in medium- to large-scale sequencing projects. The tools for EST assembly are thus essentially the same, with the limitation that when dealing with ESTs the raw data (trace files) are often not available, or are difficult to retrieve. For those interested in working from the raw data, the traces for EST sequences determined by the Washington University Genome Sequencing Center are available for downloading by FTP. It is also increasingly likely that a genomic clone matching the cluster will be found, thus aiding in the assembly process; however, because of the presence of introns, the genomic clone will have to be treated differently from the EST data.

Manual assembly software

Contig assembly can be done in essentially two ways: manual, with extensive user input (through an alignment editor) and thus more reliable consensus generation, or automatic, with the benefit of convenience but more potential for errors. A well-known tool for manual EST contig assembly is the set of GCG (Genetics Computer Group, University of Wisconsin) programs[10] used for shotgun assembly: gelstart, gelenter, gelmerge and gelassemble. These programs will create one or more contigs from a collection of sequences, and present them as editable multiple sequence alignments. The user can then manually inspect and modify the individual EST sequences and the deduced consensus. The drawbacks of this method are that the assembly and editing processes are rather time-consuming, and that some skill is required from the user to do it properly. The gap4 program, part of the Staden suite[11] distributed by the MRC Molecular Biology lab in Cambridge, has functionality similar to the GCG suite, but incorporates a wider set of methodologies for both assembly and editing of the contigs. The CRAW tools,[12] distributed by SANBI as part of their STACK_PACK suite, and commercially by Pangea Systems, are probably the most ambitious of the EST assembly tools, as they attempt additionally to distinguish splice variants and to find polymorphisms in the sequences. There are several commercial PC/Windows or Macintosh-based programs for contig assembly, of which the best-known are probably Sequencher (GeneCodes), the ContigExpress program distributed with Vector NTI (Informax), and the SeqMan module of Lasergene (DNA Star). These various manual assembly programs are well-suited for research projects where a few new gene sequences are of particular interest, but not for large-scale gene discovery projects.

Automatic assembly software

Phrap, also from Phil Green at the University of Washington, is the most widely used program for automated contig assembly in genome projects.[13] Phrap relies heavily on quality values assigned to base calls by its companion program, phred. Additionally, phrap tends to consider sequences for which there is only one read as unreliable, and thus to trim off the ends of contigs if they are not at least two ESTs confirming each other's sequence. Also, the companion alignment editor to phrap, consed, requires trace data for the edition process. As a result, the phrap/phred combination is not well suited to assemble sequence-only EST data. The cap program of Xiaoqui Huang and its more recent derivatives, cap2 and cap3,[14] do a very good job at assembling contigs for clusters of small to moderate size, but do not include an editor. When the number of cluster members goes over about 500, cap becomes too slow to be practical. TIGR also distributes an EST contig assembly program, the TIGR Assembler.[15] In the author's experience, this program is fast, but may be too stringent in the level of similarity required for assembly, thus generating an unnecessarily large number of contigs from larger clusters. For the automated

Assembled contigs

assembly of contigs from clusters of widely varying sizes, a combination of cap and phrap has been used, and an iterative assembly protocol applied to larger clusters by dividing the cluster into smaller pools, assembling these individually, and then combining the contigs. The protocol for this method, and Perl scripts implementing it, are available upon request.

The Munich Information Center for Protein Sequences (MIPS) is making available over the web a set of assembled contigs derived from human Unigene clusters, searchable by BLAST and through keyword-based methods, including the description of putative homologues.

FINDING CODING REGIONS

In order to discover new protein sequences in the EST databases, it is very helpful first to identify where potential coding regions (CDS) are located within the sequences. For obvious reasons, CDS are not annotated in the EST sequences.

CDS detection

The most common methods for searching the EST databases (see below) get around this limitation by performing searches against an automatic six-frame translation of the sequences, using, for example, the tblastn program. However, this is an extremely inefficient method:

CDS databases

the immense majority of the automatically translated sequences have no biological significance, as they consist of non-coding regions, non-coding strands and erroneous reading frames. In our experience, only about 1/32nd of the six-frame translation of an EST database corresponds to sequence from which a CDS translated in the correct frame can be deduced: five-sixths of the information is out of frame, and about five-sixths of the total sequence information is derived from non-coding regions. Therefore, the detection and extraction of true CDS from EST collections is a crucial problem.

Reducing the search space

Because of the poor quality of EST sequences, and the fact that CDS can be relatively short, the detection of open reading frames (ORF) is not a satisfactory solution to the problem. Two basic approaches to the extraction of CDS have been taken, relying either on the detection of similarities to known protein sequences or sequence motifs,[16,17] or on statistical biases in the nucleotide sequences associated with codon usage frequencies.[18,19] The second approach is more general, in that it can detect novel genes even in the absence of similarity to known proteins; however, it has to be adapted to the codon usage of the species under study. This approach has already been widely used to detect exons in genomic sequences. It has been adapted to EST analysis in a program, ESTScan, which was also designed to detect and correct sequencing errors leading to frameshifts in the CDS.[20] In its current version, ESTScan is capable of correctly predicting about 95 per cent of CDS in human ESTs, if allowed a 10 per cent 'false positive' rate. While quantitative data on its efficiency at correcting frameshift errors are not available, it was able to correct errors in all of the cases tested. Therefore, using the iterative contig assembly protocol mentioned above to generate tentative consensus sequences for EST clusters, and ESTScan to detect, correct and extract CDS, it is possible to generate a virtual protein database representing almost the full coding potential of an EST collection. As it is only a small fraction of the size of the original collection (about 1/300th for human ESTs), it can be searched efficiently, even using slow but informative methods, and with the caveat that a small fraction of true coding regions may be missing from the database.

SEARCHING THE EST DATABASES

The most common question by far asked by biologists who want to search the EST databases is: 'Are there novel genes represented among the ESTs that are related to this known gene or family of genes, or contain this known protein

domain?' There are many ways to address this question, with widely varying degrees of efficiency. The interpretation of the results is often also non-trivial and depends on how the search was performed, and on which database(s).

There are two types of queries that can be used for a search: either a sequence, or a motif descriptor. A search with a sequence asks the question: Are there any ESTs that are similar to my sequence? A search with a motif descriptor asks the question: Are there any ESTs derived from a gene encoding a protein that contains my domain? Evidently, the second question is more general, but the derivation of a usable motif descriptor is a more complex problem than the submission of a query sequence.

Searching with sequences or descriptors

Databases

EST databases

As mentioned before, the dbest database distributed by the NCBI, or the EST sections of the European Molecular Biology Laboratory (EMBL) or GenBank databases, are not split according to species of origin, and neither are they clustered or assembled. Therefore, a search against dbest is likely to produce a hit list that contains a mixture of sequences from the same species as the query (typically the query gene and paralogous genes), as well as from related species. Additionally, there can be very extensive overlaps between individual ESTs, if one or the other of the genes found by the query is abundantly represented in the database. The sorting out of a hit list produced by a search against dbest is thus often a very time-consuming, if not impossible, task, especially since the species of origin is not always properly documented in the descriptor line returned by the search program (usually BLAST).

Splitting by species

Some relief can be found by searching EST collections that have been split by species. The NCBI BLAST webpage now offers the possibility to search human, mouse or other ESTs separately. The SIB webpage also allows searches against rat and plant ESTs. An even more useful possibility is to search against EST collections that have already been clustered, ie the Unigene database. The results of such a search report the cluster numbers in addition to the accession numbers of the hits, thereby allowing a rapid differentiation among the genes that were identified by the query. The SIB BLAST pages allow searches against the human, mouse and rat Unigene collections, as well as against ESTs *not* found in Unigene, for completeness.

Contig databases

Finally, the database can consist of contigs derived from Unigene (or other) clusters. The results from such searches are obtained more quickly and are easier to interpret, but should be viewed with caution, because the contigs do not represent experimentally determined sequences, and are even more subject to artefacts than individual ESTs. The NCBI BLAST services include a webpage that allows searches against the TIGR Tentative Human Consensus (THC) database of EST contigs, and TIGR itself has a service for searching EST contigs from several species, under the denomination of unique gene indices. SANBI also produces a database of human EST contigs (STACK), clustered on the basis of their tissues of origin. STACK is searchable by BLAST on the SANBI web site. Finally, the MIPS collection of human EST contigs can be searched on their web site.

Search algorithms

The most commonly used method for searching the EST databases with a sequence query is BLAST.[21,22] This heuristic algorithm is very fast, and has been popularised by the availability of a powerful cluster of servers at the NCBI. Since the introduction of version 2 in 1997, BLAST is also able to return gapped alignments between the query and the database sequence. BLAST does, however, have a number of limitations, of which users should be aware:

BLAST

- The algorithm requires the query and the database sequence to share two small regions of significant similarity before it attempts to calculate a longer alignment; therefore, sequences whose similarity is 'diffuse', ie extends over a relatively long region without islands of stronger homology, will not be found by BLAST.

FASTA

- In order to be able to generate reasonable statistics, BLAST limits the combination of scoring matrices and gap penalty values that can be used in a protein sequence similarity search; this is not a serious limitation to the casual user, but can preclude searches where unusual scoring systems are required.

Smith and Waterman

Hardware acceleration

- The blastx, tblastn and tblastx programs, which perform a six-frame translation of either the query, or the database, or both, search only one frame at a time; therefore, sequences that contain frameshift errors (common in ESTs) may be missed.

Frameshift sensitive searches

- BLAST finds, and reports, only local alignments; if global alignment scores are needed (eg to detect whether regions of similarity between two sequences are followed by divergent regions), BLAST is not the algorithm of choice. Nevertheless, BLAST remains a robust and very useful tool for most database searches.

The FASTA algorithm[23,24] uses different heuristics from those of BLAST. It is significantly slower, and not more sensitive, for nucleotide *v.* nucleotide searches. For protein sequence comparisons, there are reports that FASTA is slightly more sensitive than BLAST at low values for the ktup parameter;[25,26] in case of doubt, it may be useful to try both methods, and compare the results. The FASTA suite has been recently completed by new programs for searching DNA databases with protein queries or vice versa.[23] These programs (fastx, fasty, tfastx and tfasty) have the advantage over their BLAST counterparts that they allow alignments to shift frames.

The most sensitive algorithm for sequence comparisons is the dynamic programming method originally described by Needleman and Wunsch[27] for global alignments, and subsequently adapted by Smith and Waterman (S-W)[28] for the calculation of local alignments. The S-W algorithm calculates an optimal alignment between the query and every sequence in the database, and reports all of the alignments that have scores above a user-selectable cut-off. Since it does not use heuristics of any kind, the S-W method is guaranteed to find all significant matches to a given query. Additionally, it can be used with arbitrary scoring systems (users beware, though!), and set-up to report either local or global alignment scores. The major limitation of the method is that it is computationally very expensive, and therefore slow on even the most powerful workstations, at least when used to search large databases. For this reason, S-W searches are usually performed on specialised hardware, designed specifically to implement dynamic programming algorithms. Vendors of such hardware include Paracel, Inc. (GeneMatcher), Compugen (Bioccelerator and BioXL) and TimeLogic Corp. (DeCypher RACE servers).

Including modules that model frameshifts, codon gaps or introns can further enhance S-W searches. Software implementing this has been developed for traditional workstations as well as for the various hardware-accelerated platforms (see eg http://www.sanger.ac.uk/Software/Wise2/).

For a cogent introduction to database search methods, it may be helpful to read the excellent summary written by Greg Schuler.[29]

Sequence queries

Nucleotide sequence queries

Since the EST databases contain nucleotide sequences, the simplest query is a nucleotide sequence itself. As has been pointed out many times, a nucleotide against nucleotide similarity search can detect only very closely related sequences. Therefore, such a search will find whether the gene from which the query was derived is represented in the database, and may also find isoforms and closely related paralogues or orthologues. As the blastn algorithm is extremely efficient and fast, a nucleotide *v.* nucleotide search is a quick and convenient way to check whether a new, unknown sequence is represented in the EST databases. It is a starting point in building a cluster that could eventually cover the sequence of the entire corresponding mRNA (*in silico* cloning and sequencing). The Italian Telethon Institute of Genetics and Medicine (TIGEM) has a public webpage to perform such a search and assembly procedure.

Detecting identities

Interactive cluster building

Repetitive sequence masking

It should be emphasised that when using sequence-based queries, and especially nucleotide sequences, care should be taken to mask out repetitive elements (eg Alu elements in human sequences) from the query. Failure to do so will result in enormous numbers of spurious hits, and make the search essentially useless. Many current BLAST servers include this possibility (usually marked as *xblast-repsim*) among their search options. Alternatives are to mask the queries with the RepeatMasker package of Arian Smit (asmit@nootka.mbt.washington.edu),

Protein sequence queries

Finding distant similarities

or to perform a BLAST against the repbase repetitive element database and to mask out the matching regions with xblast.[30]

In gene discovery applications, one is usually interested in finding new CDSs that are more distantly related to the query sequence. In this case, comparisons should be based on amino acid sequences, not nucleotides. The protein sequence of the query is normally known; as pointed out above, the CDSs represented in the ESTs are not. The simplest method for searching EST databases with a protein query is tblastn. This essentially performs an on-the-fly six-frame translation of the database, and does a standard BLAST search against this virtual protein database, with the limitations alluded to before. A faster and smarter approach is to extract likely CDSs from the EST databases first (eg using the ESTScan program), and then to perform a blastp search against this database. It has been shown that this approach produces essentially the same amount of information as a tblastn search, with the proviso that some CDSs may be missed by ESTScan and thus not found, but that others may have corrected frameshift errors and thus be 'rescued'. The most exhaustive approach is to search a six-frame translated version of the EST database using software that implements the S-W algorithm and is able to shift frames in order to extend a good alignment. Unfortunately, this type of approach is practical only if one has a hardware accelerator or dedicated server cluster at hand, or if one uses relatively small EST collections; the SIB webserver has a page that implements this method on a GeneMatcher, and allows searches against all available ESTs.

Motif-based searches

The most sensitive method by far for finding new members of known gene families in the EST databases is to use as a query a motif descriptor that embodies the information that can be extracted from a multiple alignment of all known family members or instances of the domain. There are currently two basic types of descriptors in common use:

- Position-specific scoring matrices, also known as profiles, which can come in a variety of levels of sophistication, depending on how many features of a multiple sequence alignment they have been designed to

Types of motif descriptors

represent. Subtypes of such matrices can be roughly classified as follows: (i) Short regions of gapless conserved sequence, which can be represented by simple frequency matrices; the BLOCKS[31] and PRINTS[32] databases use this type of representation. (ii) The profiles described by Gribskov *et al*.[33] introduced the possibility of gaps in the alignment, but their syntax is relatively limited. They are still in common use because the GCG software suite includes utilities for building them and using them as queries for database searches. (iii) Generalised profiles (GPs) were initially described as a more feature-rich generalisation of the Gribskov profiles.[34] They have been shown to be mathematically equivalent to hidden Markov models (HMM, see below), but include a few features (eg 'circular' profiles for the description of repeated features) that cannot be represented by HMM.[35] There is a set of public domain tools (*pftools*) for working with GP. The profile collection associated with the PROSITE database[36] is in GP format.

Hidden Markov models

Regular expressions

Motif-based searches

- HMMs were originally developed for applications unrelated to sequence analysis such as speech recognition; they can represent most of the relevant features of a sequence motif, and have been used extensively to describe protein domains and search for them. A formal description of HMM is outside of the scope of this review, but an excellent tutorial can be found in Durbin *et al*.[37] There are several public domain software packages that allow HMM construction, training and database searching. There is also an extensive collection of motifs in HMM format (Pfam)[38] that is probably the most complete set of domain descriptors available today.

It should be emphasised that the patterns used in the PROSITE database[36] are 'regular expressions' in computer science terminology. While very useful for detecting highly conserved motifs in protein sequences, they are not very satisfactory as queries for EST databases. This is because they are not amenable to a quantitative scoring system (ie partial matches will never be found) or to frameshift-tolerant searches. Motifs originally expressed as PROSITE-type patterns should probably be converted to HMM or GP format (a relatively trivial operation) before being used as queries in EST database searches.

The database requirements for motif-based searches are the same as for sequence-based ones, except that functional equivalents of tblastn and tblastx do not exist. Therefore, the EST database has to be explicitly translated into protein format before being searchable. This can be done by translating six frames independently, or by creating a 'two-frame' translation where the three frames from each strand are interleaved with one another, or by performing a CDS search and translation with ESTScan. The 'two-frame' translated databases require a special query format that can handle the threefold periodicity of the database pseudo-sequences, and the search software has to support this functionality; the pftools suite can reformat GPs, and use these to search a two-frame translated database. This method has the advantage of being amenable to frameshift-tolerant search algorithms.

Motif-based searches can take the form of a database search with a specific motif (asking the question: Does this motif appear in the database?), or a search for the appearance of known motifs in a new sequence (asking the question: Does my new sequence contain a known motif?). Gene discovery can be helped by both types of searches, but the first question is the most commonly asked. The user should be aware, however, that the proper

construction of a useful motif descriptor is a far from trivial task.

OTHER USES OF ESTS

Besides being a source of new protein sequences, EST data have also been extremely useful for a number of other purposes.

ESTs as gene tags

- While most EST sequences do not cover the full extent of the mRNAs from which they were derived, they provide 'tags' through which these mRNAs may be uniquely identified. These sequences have many important uses, among them the provision of markers for genetic and physical mapping of genomes, the association of serial analysis of gene expression (SAGE) tags with specific genes, and the design of probes for the manufacturing of microarrays for gene expression profiling.

Estimating expression levels

- In principle, the number of clones derived from the mRNA of a particular gene will be roughly proportional to the abundance of this mRNA in the tissue used to prepare the library. Hence, EST sequencing has also been used to estimate mRNA abundance in various tissues. More generally, the presence of cDNA sequences derived from a gene in libraries derived from a particular tissue provides *prima facie* evidence for the expression of that gene in that tissue.

Exon markers

- ESTs are probably the richest source of data documenting the position of exons within genes. As such, they are an indispensable complement to any genome sequencing project, since current gene and exon prediction algorithms still make an unacceptable number of errors.

PUTTING IT ALL TOGETHER

The considerations above may make it seem a daunting task to mine the EST databases for new genes of interest. It is undoubtedly true that a thorough effort will require the collaboration of a professional and well-equipped bioinformatics laboratory, with good software collections, updated local copies of the databases and high-performance hardware for database searching. Several bioinformatics companies offer such environments off the shelf, but at prices that are not affordable to most academic laboratories.

Web servers for EST work

However, thanks to the World-Wide Web, it is still possible for the average bench scientist to access the EST databases and perform useful searches, even if the processing of the results can be time-consuming. Several servers offer the possibility to perform BLAST searches on EST databases; we strongly recommend using tblastn with a protein query, and to target the search to the species or taxon of interest whenever possible. The SIB server lets users search the Unigene collections; the hits are then already assigned to clusters, thus greatly facilitating the analysis of the results and the generation of contigs if interesting clusters are found.

The SIB server also contains a webpage for submitting database searches to a GeneMatcher. In particular, it allows frameshift-aware searches of 'two-frame' translated EST databases with a protein query. The results are returned by e-mail usually within a few minutes. To our knowledge, this is the most sensitive EST search method that is publicly available today. The group of Geoff Barton at the EBI has recently put on the web an experimental server (http://circinus.ebi.ac.uk:8081/protest/), which they call protEST, that combines a tblastn search of the EST databases with an automated sorting by species and contig assembly of the hits. This should prove to be an extremely valuable resource, as it performs several of the complex tasks described above in an integrated way.

PSI-BLAST

Motif-based searches of the EST databases are best performed by specialised bioinformatics groups. This is due partly to the technical difficulties involved in deriving good-quality HMMs or profiles, and partly to the computational cost of such searches. The NCBI has recently put on the web a tool for constructing profiles from the results of BLAST searches, and then using these profiles as queries for new searches. This tool, known as PSI-BLAST,[21] has become increasingly popular because it allows non-specialists to perform true motif-based searches in a relatively simple and intuitive way. Unfortunately, only the 'traditional' protein databases are currently accessible for PSI-BLAST searching; those interested in using it for a motif-based search of a translated form of the EST databases will have to install PSI-BLAST locally.

Exploring EST data

In summary, the EST databases are an extremely rich source of new genes waiting to be discovered that has been exploited relatively little by academic scientists. A more thorough exploration of the EST data requires a good understanding of the nature and limitations of the data, and of the issues involved in searching them and sifting through the results. I hope that the present review will have clarified some of these issues. Most *in silico* gene discovery projects are just the start of a new set of experiments at the bench. Firstly, it is necessary to verify that the sequences found in the EST databases are in fact correct (they often are not…). Secondly, it is necessary to test the biological hypotheses suggested by the similarities uncovered during the database searches.

RESOURCES

Databases

dbest

EST sequences can be downloaded from the NCBI (see http://www.ncbi.nlm.nih.gov/dbEST/), or from the EBI (see http://mercury.ebi.ac.uk/dbest/dbest_index.html). Please note that dbest combines sequences from many different species, and that it is up to the end-user to sort either the database or the search results if ESTs of only one species are of interest.

Unigene

The Unigene data can be downloaded from the NCBI either as lists of accession numbers making up a cluster, or as the actual sequences sorted by cluster. See http://www.ncbi.nlm.nih.gov/UniGene/.

Pfam

The Pfam database of protein domains in HMM format can be downloaded from several locations. See http://pfam.wustl.edu/.

PROSITE

The PROSITE collection of protein domains in GP format can be downloaded from the Swiss Institute of Bioinformatics, at ftp://ftp.isrec.isb-sib.ch/pub/sib-isrec/profiles/. The same server also has the Pfam collection converted to GP format.

BLOCKS

The BLOCKS database of locally conserved protein sequences is available from the Fred Hutchison Center for Cancer Research. See http://blocks.fhcrc.org/blocks/blocks_release.html.

PRINTS

Terri Attwood's PRINTS database can be accessed at University College London. See http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/.

Software

BLAST

The BLAST programs are available as executables or as source code from the NCBI. See http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html. We have developed a distributed client/server environment that allows the

dispatching of BLAST jobs to multiple servers and the implementation of database 'farms' for simultaneous searches. Contact Christian.Iseli@licr.org if you are interested in obtaining this software.

FASTA and ssearch
The FASTA and ssearch (an implementation of the Smith and Waterman algorithm) programs are available from Bill Pearson at the University of Virginia. See http://www.cs.virginia.edu/brochure/profs/pearson.html.

Wise2
The Wise2 package of Ewan Birney at the Sanger Centre combines many interesting database search algorithms, in particular for aligning protein sequences with DNA. See http://www.sanger.ac.uk/Software/Wise2/.

HMMER
The HMMER package developed by Sean Eddy (Washington University) is a comprehensive set of tools for generating, calibrating and searching with HMMs. See http://hmmer.wustl.edu/.

SAM
The SAM package of Richard Hughey and Anders Krogh (UC Santa Cruz) is very similar to HMMER in its functionalities, but uses a different HMM file format. Contact sam-info@cse.ucsc.edu for information about SAM.

Pftools
The pftools package is a comprehensive set of tools developed by Philipp Bucher for working with GPs. It can be downloaded from the Swiss Institute of Bioinformatics at ftp://ftp.isrec.isb-sib.ch/pub/sib-isrec/pftools/.

Phred and phrap
The phred base-calling program and the phrap contig assembler can be obtained from Phil Green (phg@u.washington.edu) at the University of Washington.

Cap, cap2 and cap3
The cap series of contig assembly programs are available from Xiaoqiu Huang (Michigan State University) at huang@mtu.edu. See http://genome.cs.mtu.edu/cap/cap3.html.

GCG
The GCG suite of programs, originally developed by the Genetics Computer Group at the University of Wisconsin, has been commercial software for many years. It is accessible through the computer centres of many academic institutions, and through the national EMBnet nodes of almost all European countries. See http://www. gcg.com.

Staden package
The Staden package is a complete program suite for contig assembly from trace files, with many basic sequence analysis functions thrown in. It now exists for both Unix and Windows NT, and is free to academic users. See http://www.mrc-lmb.cam.ac.uk/pubseq/ for more information.

Webservers

National Center for Biotechnology Information: www.ncbi.nlm.nih.gov
Home of the dbest and Unigene databases, and of the BLAST search services. Excellent source of information about EST databases. The NCBI BLAST pages allow searches of human, mouse and other ESTs.

European Bioinformatics Institute: www.ebi.ac.uk
Home of the EMBL sequence database. Offers more search algorithms than the NCBI server (eg FASTA, Smith–Waterman), but only against the entire EST database. The tfastx and tfasty algorithms are particularly useful, as they

accommodate frameshifts in the EST database sequences. The EBI webserver also offers a number of resources for EST analysis, clustering and assembly, at http://corba.ebi.ac.uk/EST/.

Swiss Institute of Bioinformatics: www.ch.embnet.org
Home of the GP methodology and the PROSITE profile collection. Offers BLAST searches against many EST collections, including the Unigene clusters, and frame-tolerant, hardware-accelerated Smith–Waterman searches against translated EST collections. Also offers coding region detection (ESTScan) and detection of protein domains in EST-quality DNA sequences (pframescan).

Sanger Centre: www.sanger.ac.uk
Primarily a genome centre, but also has some interesting tools for EST analysis, in particular Wise2 (protein *v.* DNA alignments), protEST (protein *v.* EST database searches), and searches of DNA sequences against the Pfam database.

Washington University Genome Center: http://genome.wustl.edu
Source of much of the EST sequences available today, and repository for primary information about the clones and their sequences. This is where the original trace files for many EST sequences can be downloaded.

South African National Bioinformatics Institute: www.sanbi.ac.za
Home of the STACK database of EST contigs. Offers BLAST searches against STACK, and links of the results to dbest.

The Institute of Genome Research: www.tigr.org
TIGR, the institution where EST sequencing started, has recently created Unique Gene Indices of EST contigs for many species, which can be searched using BLAST.

Italian Telethon Institute of Genetics and Medicine: http://www.tigem.it/
TIGEM offers a semi-automated EST search and assembly service, as well as a comprehensive collection of links to other webservers that host EST databases and search engines.

Munich Information Center for Protein Sequences (MIPS): http://www.mips.biochem.mpg.de
MIPS has produced human EST assemblies from Unigene clusters; these can be searched by BLAST or through the SRS indexing system.

Human Genome Mapping Project Resource Centre: http://www.hgmp.mrc.ac.uk
This site is working on the construction of the HUGEN minimal gene set in collaboration with the Sanger Centre.

Biocomputing Service Group, German Cancer Research Center: http://genome.dkfz-heidelberg.de
Both of these sites offer EST clustering and assembly services, starting with a user-defined query. Both require that users be registered (and paid up) to access these services.

References

1. Gerhold, D. and Caskey, C. T. (1996), 'It's the genes! EST access to human genome content'. *Bioessays*, Vol. 18(12), pp. 973–981.
2. Adams, M. D. *et al.* (1991), 'Complementary DNA sequencing: expressed sequence tags and human genome project', *Science*, Vol. 252(5013), pp. 1651–1656.
3. Ewing, B. *et al.* (1998), 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment', *Genome Res.*, Vol. 8(3), pp. 175–185.
4. Bonaldo, M. F., Lennon, G. and Soares, M. B. (1996), 'Normalization and subtraction: two approaches to facilitate gene discovery', *Genome Res.*, Vol. 6(9), pp. 791–806.

5. Gautheret, D. *et al.* (1998), 'Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering', *PCR Methods Appl.*, Vol. 8(5), pp. 524–530.
6. Parsons, J. D. (1995), 'Improved tools for DNA comparison and clustering', *Comput. Appl. Biosci.*, Vol. 11(6), pp. 603–613.
7. Gill, R. W. *et al.* (1997), 'A new dynamic tool to perform assembly of expressed sequence tags (ESTs)', *Comput. Appl. Biosci.*, Vol. 13(4), pp. 453–457.
8. Schuler, G. D. (1997), 'Pieces of the puzzle: expressed sequence tags and the catalog of human genes', *J. Mol. Med.*, Vol. 75(10), pp. 694–698.
9. Adams, M. D. *et al.* (1995), 'Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence', *Nature*, Vol. 377(6547 Suppl.), pp. 173–174.
10. Devereux, J., Haeberli, P. and Smithies, O. (1984), 'A comprehensive set of sequence analysis programs for the VAX', *Nucleic Acids Res.*, Vol. 12, pp. 387–395.
11. Staden, R. (1996), 'The Staden sequence analysis package', *Mol. Biotechnol.*, Vol. 5(3), pp. 233–241.
12. Burke, J. *et al.* (1998), 'Alternative gene form discovery and candidate gene selection from gene indexing projects', *Genome Res.*, Vol. 8(3), pp. 276–290.
13. Gordon, D., Abajian, C. and Green, P. (1998), 'Consed: a graphical tool for sequence finishing', *Genome Res.*, Vol. 8(3), pp. 195–202.
14. Huang, X. (1996), 'An improved sequence assembly program', *Genomics*, Vol. 33(1), pp. 21–31.
15. Fleischmann, R. D. *et al.* (1995), 'Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd', *Science*, Vol. 269(5223), pp. 496–512.
16. Brown, N. P., Sander, C. and Bork, P. (1998), 'Frame: detection of genomic sequencing errors', *Bioinformatics*, Vol. 14(4), pp. 367–371.
17. Birney, E., Thompson, J. D. and Gibson, T. J. (1996), 'PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames', *Nucleic Acids Res.*, Vol. 24(14), pp. 2730–2739.
18. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.*, Vol. 268(1), pp. 78–94.
19. Lukashin, A. V. and Borodovsky, M. (1998), 'GeneMark.hmm: new solutions for gene finding', *Nucleic Acids Res.*, Vol. 26(4), pp. 1107–1115.
20. Iseli, C., Jongeneel, C. V. and Bucher, P. (1999), 'ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences', 'ISMB 99', Vol. 7, Lengauer, T. *et al.*, Eds, AAAI Press, Menlo Park, CA. pp. 138–147.
21. Altschul, S. F. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.
22. Altschul, S. F. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215, pp. 403–410.
23. Pearson, W. R. *et al.* (1997), 'Comparison of DNA sequences with protein sequences', *Genomics*, Vol. 46(1), pp. 24–36.
24. Pearson, W. R. (1990), 'Rapid and sensitive sequence comparison with FASTP and FASTA', *Meth. Enzymol.*, Vol. 183, pp. 63–98.
25. Pearson, W. R. (1991), 'Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms', *Genomics*, Vol. 11, pp. 635–650.
26. Anderson, I. and Brass, A. (1998), 'Searching DNA databases for similarities to DNA sequences: when is a match significant?', *Bioinformatics*, Vol. 14(4), pp. 349–356.
27. Needleman, S. B. and Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *J. Mol. Biol.*, Vol. 48, pp. 443–453.
28. Smith, T. F. and Waterman, M. S. (1981), 'Comparison of bio-sequences', *Adv. Appl. Math.*, Vol. 2, pp. 482–489.
29. Schuler, G. D. (1998), 'Sequence alignment and database searching', in 'Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins', Baxevanis, A. D. and Ouellette, B. F. F, Eds., Wiley-Liss, New York, pp. 145–171.
30. Claverie, J.-M. and States, J. D. (1993), 'Information enhancement methods for large scale sequence analysis', *Comput. Chem.*, Vol. 17, pp. 191–201.
31. Henikoff, S., Henikoff, J. G. and Pietrokovski, S. (1999), 'Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations', *Bioinformatics*, Vol. 15(6), pp. 471–479.
32. Attwood, T. K. *et al.* (1997), 'The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology', *J. Chem. Inf. Comput. Sci.*, Vol. 37(3), pp. 417–424.

33. Gribskov, M., Luethy, R. and Eisenberg, D. (1990), 'Profile analysis', *Meth. Enzymol.*, Vol. 183, pp. 146–159.

34. Bucher, P. and Bairoch, A. (1994), 'A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation', in 'ISMB-94; Proceedings Second International Conference on Intelligent Systems for Molecular Biology', Altman, R., Brutlay, D., Karp, P., Lathrop, R. and Searls, D., Eds, AAAI Press, Menlo Park, pp. 53–61.

35. Bucher, P. and Hofmann, K. (1996), 'A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system', 'ISMB', Vol. 4, Lengauer, T. *et al.*, eds, AAAI Press, Menlo Park, CA, pp. 44–51.

36. Hofmann, K. *et al.* (1999), 'The PROSITE database, its status in 1999', *Nucleic Acids Res.*, Vol. 27(1), pp. 215–219.

37. Durbin, R. *et al.* (1998), 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press, Cambridge.

38. Sonnhammer, E. L. *et al.* (1998), 'Pfam: multiple sequence alignments and HMM-profiles of protein domains', *Nucleic Acids Res.*, Vol. 26(1), pp. 320–322.

APPENDIX

Table 1: EST database distribution resources

Database	Distributor*	Type	Species	Access
dbEST, ESTs in GenBank	NCBI	Raw data	All (unsorted)	FTP, keyword search, BLAST
ESTs in EMBL	EBI	Raw data	All (unsorted)	FTP, keyword search, FASTA
Unigene	NCBI	Clusters	Human, mouse, rat	FTP, keyword search
Gene indices	TIGR	Contigs	Human, mouse, rat, zebrafish, *Drosophila*, Arabidopsis, rice, soybean, tomato	BLAST, FTP (need licence)
STACK	SANBI	Contigs, clustered by tissue	Human only	BLAST, FTP (need licence)
LifeSeq†	Incyte	Clusters and contigs	Human only	By contract only

* NCBI, National Center for Biotechnology Information, Bethesda, MD, USA; EBI, European Bioinformatics Institute, Hinxton, UK; TIGR, The Institute for Genomic Research, Rockville, MD, USA; SANBI, South African National Bioinformatics Institute, Cape Town, South Africa; Incyte, Incyte Pharmaceuticals, Inc., Palo Alto, CA, USA.

† Incyte has produced its own collection of EST sequences, which are not in the public domain. The LifeSeq products come in versions that do or do not include these proprietary data.

Table 2: EST database search resources

Website	Databases	Search methods	Remarks
NCBI	dbEST, mouse, human, others	BLAST	Fastest search site available, but often overloaded; supports BLAST only
EBI	ESTs from EMBL	BLAST (NCBI & WU) FASTA	Includes tblastn and tfastx Javascript should be enabled
SIB	Human, mouse rat, *Drosophila*, plants, others, Unigene human, rat, mouse	BLAST, S-W with frameshifts	Only site where framesearches against EST databases are available
Infobiogen	ESTs from Genbank, dbEST	BLAST, FASTA	Includes tblastn and tfastx
SANBI	STACK, predicted proteins	BLAST	Includes tblastn, and blastp against proteins predicted from STACK contigs
MIPS	Unigene contigs and predicted proteins	BLAST	Documents assembly of the EST contigs
TIGR	Gene indices (see Table 1)	BLAST	Includes both blastn and tblastn. Results documents assembly and are linked to ATCC clone numbers
TIGEM	ESTs from EMBL, Unigene	BLAST	Includes an EST clustering and contig assembly utility

On most servers, tblastn searches take too long for interactive retrieval of the results. All servers offer the possibility to return them by e-mail (but without hyperlinks).

The URLs of the webservers are given in the Resources list, except for Infobiogen (France): http://www.infobiogen.fr